Analyzing the quality of CDC and JHU COVID-19 Daily State Data

Kushajveer Singh
University of Georgia
Athens, Georgia
ks56866@uga.edu

Abstract

Over the course of the last two years, researchers have been trying to predict the spread of the COVID-19 pandemic through data using various machine learning and deep learning models. In order to aid the researchers, two organizations CDC *and* CSSE-JHU *have put in tremendous effort to make the data readily available for fair use. Although there have been many papers in the last two years that used these datasets, no studies have been done on the quality of the datasets, especially the number of daily deaths reported by both CDC and JHU as this is the main feature that is predicted by the various time series models. We find that there are some huge discrepancies between the number of deaths reported by the CDC and JHU datasets for multiple states of the United States, and in order to help people visualize these discrepancies we also release an interactive web demo that people can use to see the differences in the number of deaths reported by CDC and JHU for a given date and how the CDC and JHU data looked around that period. We find some striking patterns, possibly due to human errors in the reporting of the data, that seem to be hard for models to generalize.*

1. Introduction

The COVID-19 pandemic caused by the SARS-CoV-2 virus has led to over 6.65 million deaths worldwide and around 1.09 million deaths in the United States while infecting over 649 million people globally and 99.1 million people in the United States[1]. The United States has faced one of the worst consequences of the global pandemic caused by COVID-19, with over 1 million reported deaths [2]. Due to this reason, it is of utmost importance that we learn from this experience and build techniques that can help up analyze pandemics like this in the future.

Since the COVID-19 pandemic occurred in the digital era we have significantly more data that can be leveraged to provide us with useful information, than the previous pandemics. One way of using the data is to analyze it for any patterns and then in the future, we can repeat the techniques for any new variant of the virus. Knowing how the disease is spreading and how it will spread in the future is important, as this information can be used by government agencies to better prepare for the pandemic. Time-series forecasting provides a way to predict the number of deaths in the future and this information can be used by the government and healthcare providers to better allocate medical resources and ensure public safety.

When making predictions about future deaths, a lot of factors need to be accounted for like hospitalization rate, topological factors like distance and population density, the testing rate, and much more. Further, the spread of COVID-19 is significantly different in different states of the United States. California has the highest number of cases among all the other states at 11.6 million, while Mississippi has the highest number of deaths per 100,000 people at 439 and California at 248 deaths per 100k people [2]. Seeing the complex nature of the problem and how many things need to be accounted for to make a realistic prediction for the future deaths of the pandemic, using statistical models like

[1]The cases and deaths are reported from National Center for Health Statistics https://www.cdc.gov/nchs.

[2]Data reported from https://www.statista.com/page/covid-19-coronavirus

ARIMA, SARIMA, SARIMAX [4] seems to be the better option. Further, to better account for the relationship between all the features Deep Learning techniques like LSTM [8], Graph Neural Networks [10], recurrent neural networks [7] and Transformer based Graph Convolutional Neural Network models [14] can be used.

Good data is extremely important for training good machine learning models, to the point where people can spend up to 80 percent of the time on data preparation only. Since data collection and data, preparation takes such a long time various organizations have stepped up to provide this data to the researchers for use, thus allowing the researchers to focus on the modeling part. Among these organizations, the most widely used datasets are *CDC COVID Data Tracker dataset* and *CSSE-JHU COVID-19 dataset* for the state-level data. Since these organizations have been collecting and aggregating medical data, independently of each other such as the number of infected individuals, number of deaths, number of hospitalized patients, number of patients in the Intensive Care Unit, number of individuals vaccinated, etc, reporting errors are expected. There are multiple reasons for these discrepancies including reporting errors, miscalculations, systematic biases, and rollbacks in various data sources. Therefore, it is really important that we ensure the dataset is of good quality.

In this paper, we focus specifically on the number of deaths as this is the metric we are trying to predict in the time-series models. We find that the number of deaths in the CDC and JHU data is not similar and even when we take multiple days moving averages there are huge discrepancies between the datasets. So the goal of this paper is to explore the two datasets in detail and do exploratory data analysis to identify the discrepancies. Further, we find that there are some states that have much worse disagreement between the number of deaths in the CDC and JHU datasets. Although there have been many research papers presented in the last two years, very few have focused [3] on the quality of the datasets and exploring how the number of deaths reported by the CDC COVID Data Tracker and CSSE-JHU COVID-19 dataset differ.

2. Datasets

Various organizations have released COVID-19 data sources since the start of the pandemic, which can be used to get the number of cases, daily deaths, hospitalization rate, and vaccination rate. The two most widely used datasets are *CDC COVID Data Tracker* [1] and *CSSE-JHU COVID-19 Dataset* [6]. In the next subsection, we discuss these two datasets in detail and the various data processing techniques that were applied to get the final dataset.

2.1. CDC COVID Dataset

CDC is considered the official dataset for the COVID-19 pandemic. CDC has been collecting the data since the beginning of January 22, 2020 (as shown in Tab. 1. Although there is some evidence that the patient zero was around November 17, 2019 [15], we would consider the CDC start date to be official as worldwide testing and reporting also started around that time, and even looking at the data for the various states of the United States, we can see that it took an average of 56 days for the first death to be reported, with a maximum time of 74 days in the Virgin Islands and a minimum time of 36 days in Alabama. This information is crucial to align the CDC and JHU datasets, as we need to drop the start dates from the CDC dataset, as shown in Tab. 1.

CDC depends upon the voluntary reporting of the data including hospitalization rate, the number of cases, and the number of deaths from the local, state, and territorial departments. The government bodies further receive the data from the local hospitals, healthcare providers, and laboratories as per the local reporting laws, and this creates some discrepancies in the time frame in which the data needs to be reported. Most of the time healthcare providers are required to provide the data within 7 days of reporting, but over the course of weekends and holidays the data may not be reported and as a result, the healthcare providers are required to provide the data on the next day with appropriate time stamps. Also, some of the data is even reported by phone or by hand and this further adds human error in the data collection process. For reporting of deaths and getting the weekly counts, CDC depends upon the states to submit death certificates and as a result CDC has to continuously revise previously

Dataset	Start Date	End Date	Rows per state
CDC	Jan 22, 2020	Oct 18, 2022	1001
JHU	Apr 12, 2020	Nov 17, 2022	950

Table 1. Overview of the CDC and JHU datasets. Although CDC started reporting the data on Jan 22, 2020, on average the first 56 days did see any reporting of deaths. CDC stopped reporting the state daily data on Oct 18, 2022, but JHU is still releasing the daily data. The table shows the data as of Nov 17, 2022.

State	Diff
New York	7365
Missouri	2441
Tennessee	2047
New Jersey	1743
Oklahoma	1713
Florida	1269

Table 2. Top 6 states with the largest maximum absolute difference between the daily deaths reported by CDC and JHU.

released data as states report at different rates, and although 80% of deaths are electronically processed, most deaths are processed by a person which takes an average of 7 days[3].

We collect the CDC data directly from their homepage. To make the CDC dataset similar to the JHU dataset we need to drop some smaller territories from the dataset which include American Samoa, Marshall Islands, New York City (as we are state analysis and this information is not available in the JHU dataset), Guam, Northern Mariana Islands, Federated States of Micronesia, and Palau. CDC provides daily deaths under the column *new_death* and we can use it to directly compute the moving average of 3 and 7 days. Further, there are a lot of missing values in the dataset for the columns conf_cases, prob_cases, conf_death, prob_death and consent_cases and we find that using these columns for training any downstream models to be not useful and reliable.

2.2. CSSE-JHU COVID-19 Dataset

Center for Systems Science and Engineering at the Johns Hopkins University (CSSE-JHU) provides COVID-19 daily state data for the United States starting from April 12, 2020, Tab. 1. Unlike, CDC which discontinued the daily state data collection process on October 18, 2022, JHU is still providing the daily data. Although, for this paper, we only use the data till November 17, 2022. A lot of prior research papers have focused on using this dataset exclusively, including [16], [12], [11] where authors trained Long Short Term Memory (LSTM) [8], Gated Recurrent Networks (GRUs) [5] and various Variational Auto-Encoder (VAE) [9] models. Due to the prevalence of the JHU dataset, it is important we point out the errors in the dataset which can further help researchers identify issues relating to poor performance for certain data ranges for a particular state.

To prepare the JHU dataset we use the CSV files from the official CSSE-JHU github repo [4]. To make the dataset consistent with CDC data some states and smaller territories are dropped from the table, which includes American Samoa, Diamond Princess, Grand Princess, Guam, Northern Mariana Islands and Recovered (which is a unique value present in the state column). As JHU does not provide daily death information directly, we use the cumulative death column *Deaths* to compute daily deaths in *daily_deaths* column, and finally, we compute the rolling average of daily deaths with a window of size 3 and 7. Unlike the CDC dataset, the JHU dataset only contains 2 missing columns with missing values, specifically People_Tested and Mortality_Rate.

2.3. Scalation COVID-19 Dataset

To provide easier access to the datasets used in the paper and for increased reproducibility, you can find all the data and code in Scalation [13] COVID-19 dataset[5]. The datasets used in the paper are located in COVID-State/data_analysis/data along with the scripts in the same folder which are used to create the dataset.

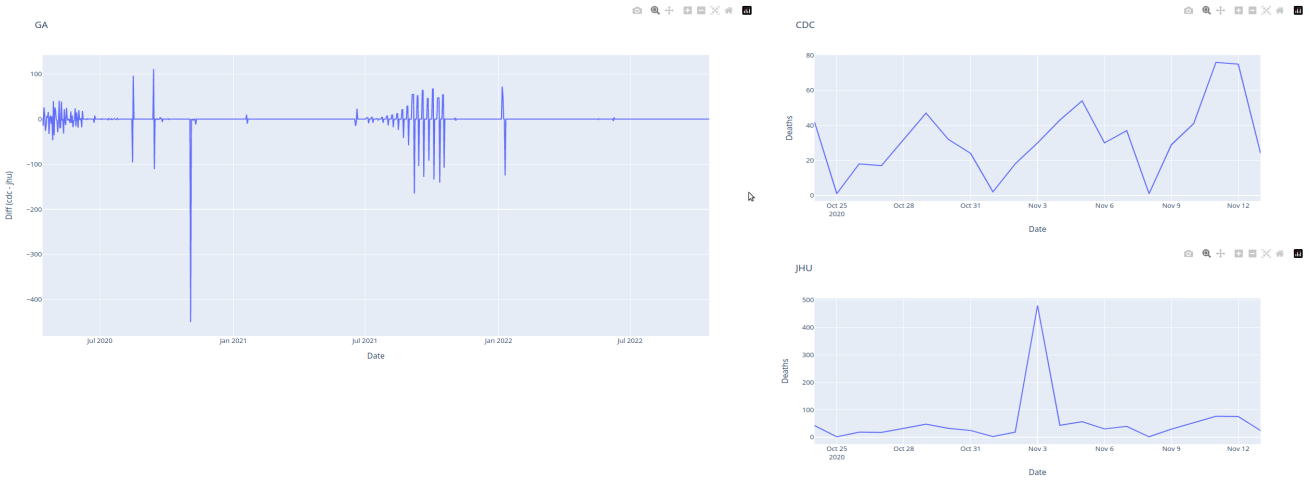
3. Data Quality

Multivariate-time series models that would use the CDC or JHU data would predict the number of daily

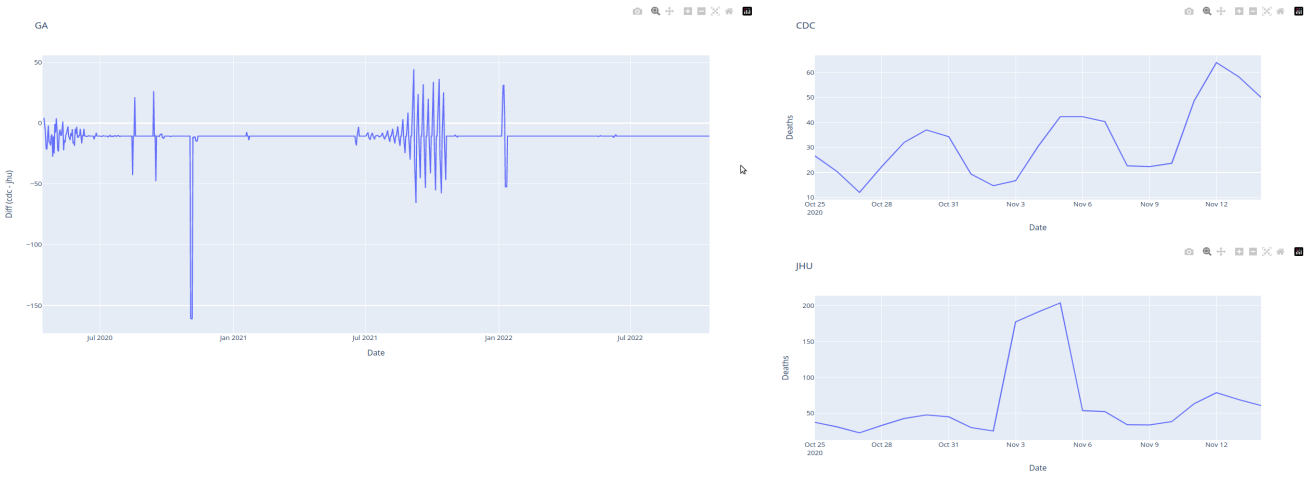
[3] https://www.cdc.gov/nchs/nvss/vsrr/covid_weekly/index.htm

[4] https://github.com/CSSEGISandData/COVID-19

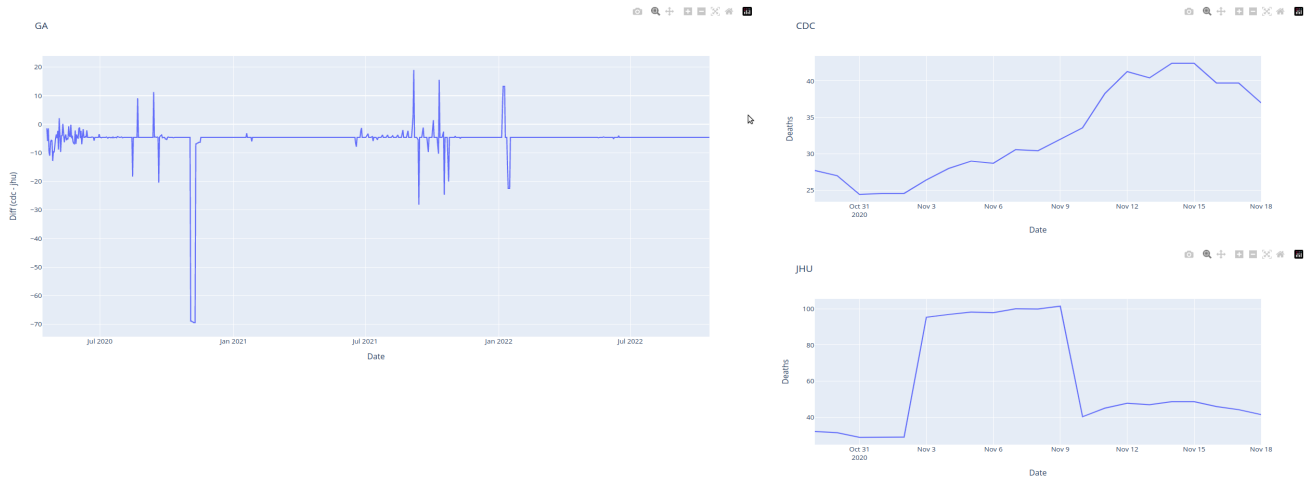
[5] https://github.com/scalation/data



(a) Daily deaths in GA

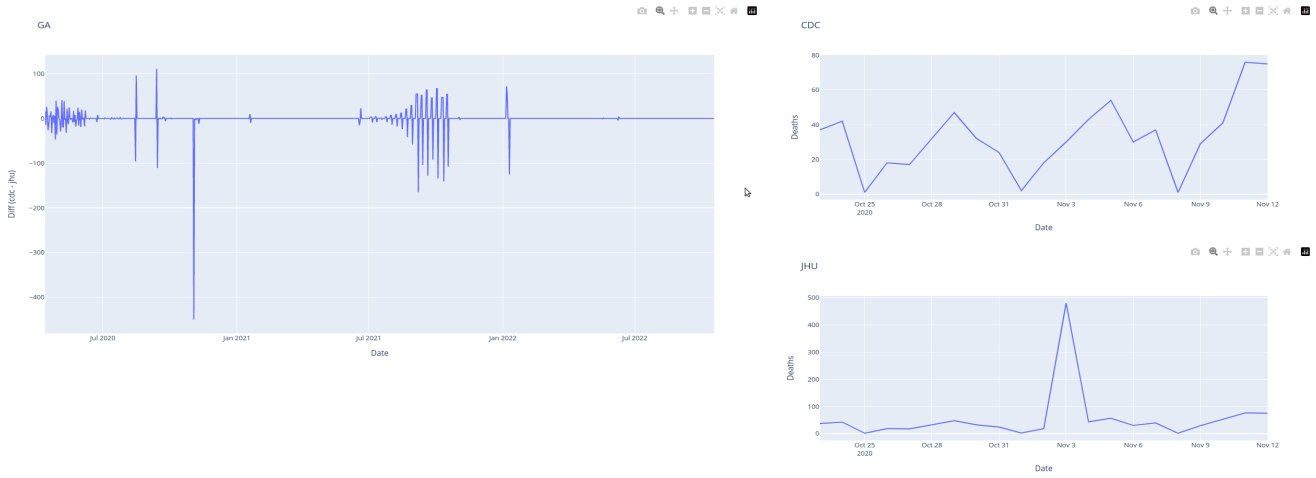


(b) Moving average of daily deaths with window size 3 in GA

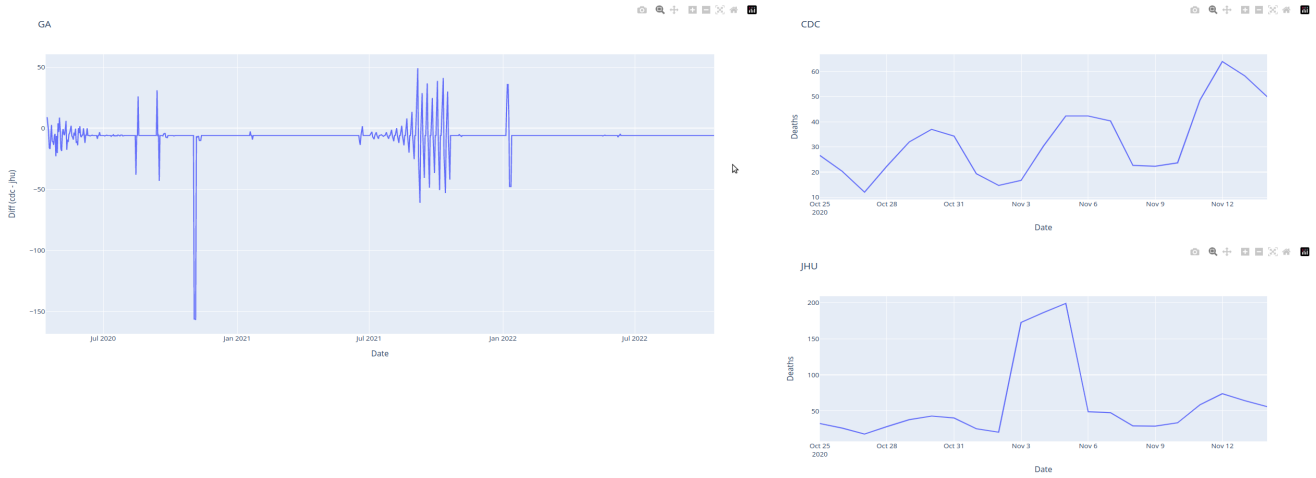


(c) Moving average of daily deaths with window size 7 in GA

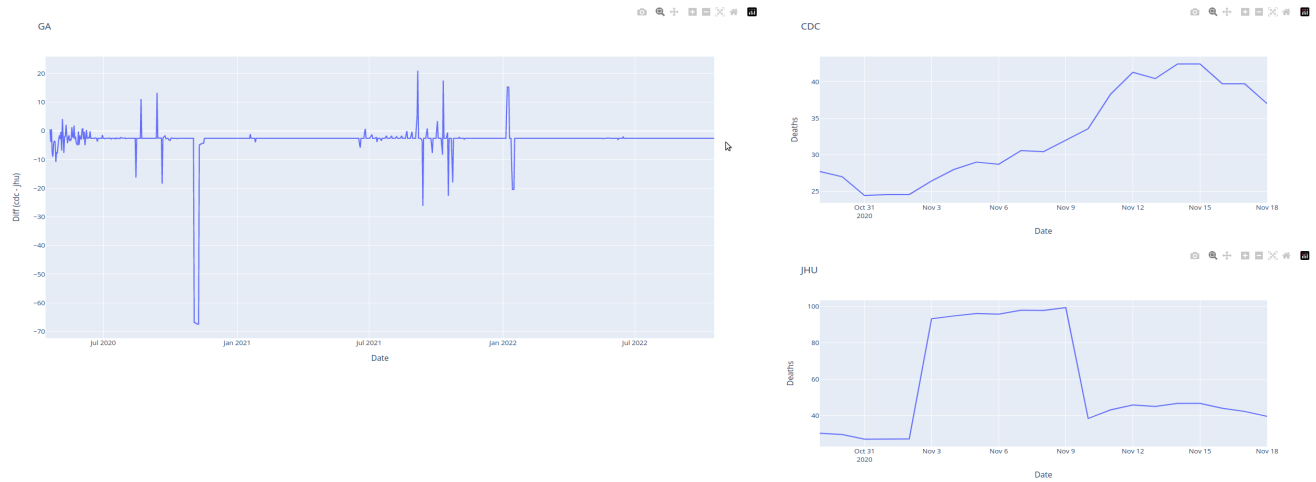
Figure 1. For the top subfigure, on the left the difference between the number of deaths of CDC and JHU is shown. On top-right, CDC data is shown for the point at which the difference between CDC and JHU is maximum, with a window size of 10. On the bottom right, JHU data is shown for the point of maximum difference. Using this visualization we can easily see that there is a huge discrepancy between the number of deaths reported by CDC and JHU.



(a) Daily deaths in GA

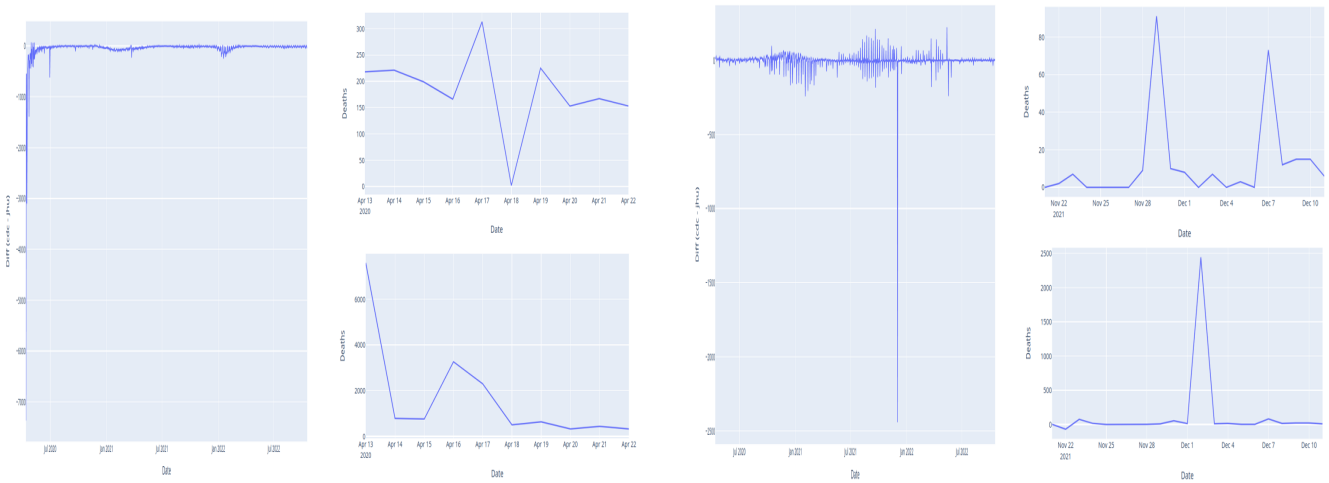


(b) Moving average of daily deaths with window size 3 in GA



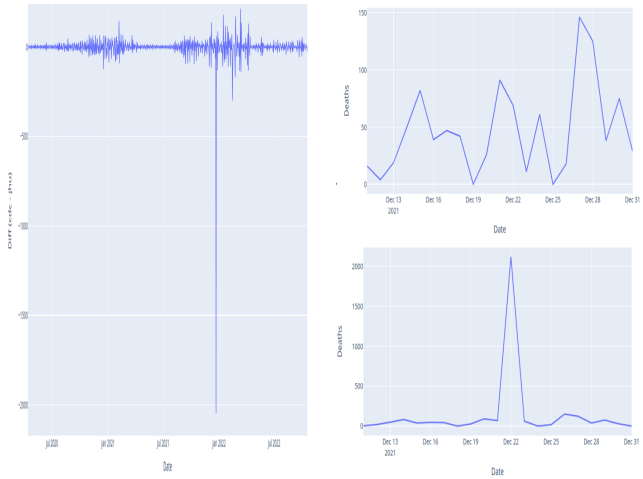
(c) Moving average of daily deaths with window size 7 in GA

Figure 2. This is similar to Fig. 1, except we are using the total deaths from CDC to start the JHU cumulative deaths. This is done because many papers consider CDC data to be official, and thus it makes sense to start the JHU using the value from CDC data.

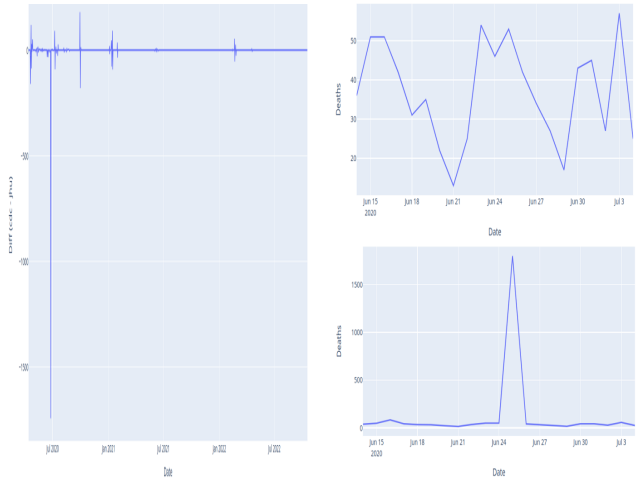


(a) New York

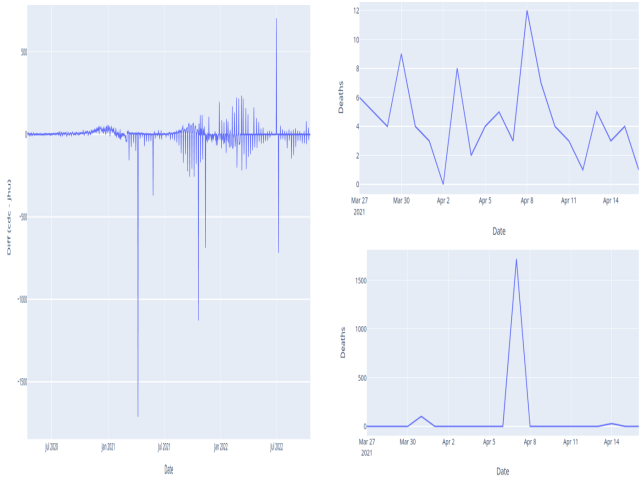
(b) Missouri



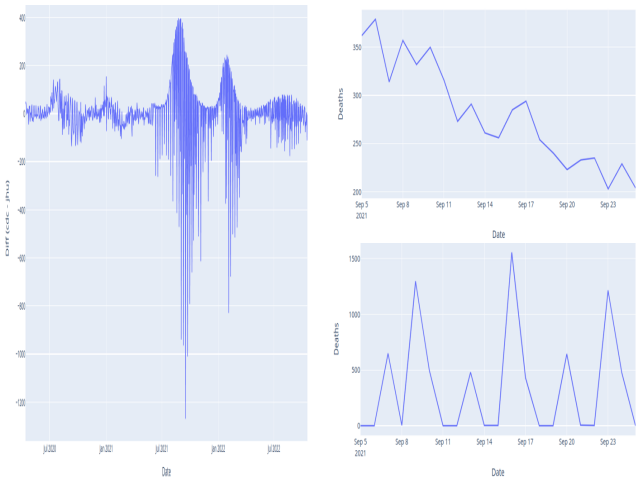
(c) Tennessee



(d) New Jersey



(e) Oklahoma



(f) Oklahoma

Figure 3. Similar to Fig. 1, the top-6 states with the largest absolute maximum difference between the reported daily deaths between CDC and JHU are shown. One pattern that emerges from these figures is that majority of these errors are caused by poor reporting by JHU data for some days, where we are able to see some cyclic patterns. Researchers trying to train machine learning models on this kind of data need to be careful, on how to deal with these outliers as these can lead to significantly poor results.

deaths for a given window in the future. But as discussed in the previous section, there are a lot of reporting errors including entering data by phone and by hand. To mitigate these issues, it is recommended that instead of using the daily deaths to train the model, using a rolling average of 7 days is much more beneficial.

To visualize the discrepancies between the daily death data of CDC and JHU, we have created a Flask app[6] that provides an interactive visualization to see the CDC and JHU data for any date and the neighboring data with a window size of 10. You can use this tool to further visualize daily deaths, rolling averages of 3 days and 7 days.

Fig. 1 shows the discrepancy between the CDC and JHU datasets. Here we visualize the difference between the deaths reported by CDC and JHU for the state of GA and then on the right we visualize how the CDC and JHU data look around the date with a maximum absolute difference, and as we can see the data reported by CDC and JHU is significantly different from each other. To further look into the issue, Fig. 2 shows a similar visualization but we use the total deaths reported by CDC as the starting value for JHU total deaths. This is done because a lot of papers consider the CDC data to be ground truth and as we can see from Fig. 2 there are still huge discrepancies between the datasets. Even using a rolling average is not sufficient to mitigate the reporting errors.

Tab. 2 shows the top-6 states with the largest absolute maximum difference between the daily deaths reported by CDC and JHU. Further looking into Fig. 3 we can see some shocking patterns that emerge for the daily deaths reported by the CDC and JHU. We see some striking patterns in Fig. 3 where we observe that majority of the errors are caused by poor reporting in JHU data. As we can see some significant outliers and researchers trying to train machine learning models on JHU data need to be careful about these discrepancies as it really hard to train a machine learning model that can mimic the cyclic nature as shown in the state of Florida in the JHU dataset.

[6] COVID-State/data_analysis/interactive_plot.py can be used to run 'python app.py' and it would open a web server at https://127.0.0.1:8050/

4. Conclusion

In this paper, we looked at the quality of COVID-19 daily state data for the United State provided by two organizations CDC and JHU, which have been significantly used by researchers in the past two years to train various machine learning models for multivariate time series forecasting or for learning about the spread of the pandemic. We find that there is a significant discrepancy between the daily deaths reported by the CDC and JHU datasets. One major contributor to this discrepancy is the errors added due to humans in the loop as some portion of the data is still collected by phone or from handwritten paper. We also check for discrepancies in case we assume CDC to be the ground-truth data, and replacing the JHU total deaths with CDC total deaths for the first date still leads to similar errors. Hence, it is useful for researchers trying to train various machine learning and deep learning models for multivariate time series forecasting to be familiar with these discrepancies, as we observed that there are some patterns in the JHU dataset that appear to be very hard for a model to generalize and treating them as outliers can also be considered an option.

References

[1] Cdc covid data tracker, 2020. https://covid.cdc.gov/covid-data-tracker. 2

[2] Who covid-19 dashboard, 2021. https://covid19.who.int/. 1

[3] Beth Blauer. The state of state-level breakthrough case reporting. 2021. 2

[4] George.E.P. Box and Gwilym M. Jenkins. *Time Series Analysis: Forecasting and Control*. Holden-Day, 1976. 2

[5] Junyoung Chung, Çaglar Gülçehre, KyungHyun Cho, and Yoshua Bengio. Empirical evaluation of gated recurrent neural networks on sequence modeling. *CoRR*, abs/1412.3555, 2014. 3

[6] Ensheng Dong, Hongru Du, and Lauren Gardner. An interactive web-based dashboard to track covid-19 in real time. *The Lancet Infectious Diseases*, 20(5):533–534, 2020. 2

[7] Hansika Hewamalage, Christoph Bergmeir, and Kasun Bandara. Recurrent neural networks for time series forecasting: Current status and future directions. *International Journal of Forecasting*, 37(1):388–427, 2021. 2

[8] Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural computation*, 9(8):1735–1780, 1997. 2, 3

[9] Diederik P Kingma and Max Welling. Auto-encoding variational bayes, 2013. 3

[10] Thomas N. Kipf and Max Welling. Semi-supervised classification with graph convolutional networks, 2016. 2

[11] Naresh Kumar and Seba Susan. Covid-19 pandemic prediction using time series forecasting models. In *2020 11th international conference on computing, communication and networking technologies (ICCCNT)*, pages 1–7. IEEE, 2020. 3

[12] Mohsen Maleki, Mohammad Reza Mahmoudi, Darren Wraith, and Kim-Hung Pho. Time series modelling to forecast the confirmed and recovered cases of covid-19. *Travel medicine and infectious disease*, 37:101742, 2020. 3

[13] John A Miller. Introduction to computational data science using scalation. 2022. 3

[14] Yunsheng Shi, Zhengjie Huang, Shikun Feng, Hui Zhong, Wenjin Wang, and Yu Sun. Masked label prediction: Unified message passing model for semi-supervised classification, 2020. 2

[15] Sedlák V Tichopád A, Pecen L. Could the new coronavirus have infected humans prior november 2019? *PLoS One. 2021;16(8):e0248255*, 2021 Aug 19. 2

[16] Abdelhafid Zeroual, Fouzi Harrou, Abdelkader Dairi, and Ying Sun. Deep learning methods for forecasting covid-19 time-series data: A comparative study. *Chaos, Solitons & Fractals*, 140:110121, 2020. 3

Appendix

1. Knowledge Graph

Making predictions for the spread of the COVID-19 pandemic is a complex task, which can benefit from the information available from multiple sources. To integrate this information into the multi-variate time series models, we can build a knowledge graph using all the information and use the knowledge graph to create embeddings that can be used with the time series models. Using knowledge graphs can improve the situational awareness of the models as we incorporate more information from the real-world and thus producing more accurate and robust forecasts.

1.1. Building knowledge graph

Building the knowledge graph consists of defining a schema and updating the knowledge graph dynamically. For building the schema we need to first identify how the knowledge graph embeddings are going to integrate with the time-series models. For example, are we going to use the embeddings of ”date” or ”state” or even multiple variables when integrating with the time-series models? These variables will act as a bridge between the information stored in the knowledge graph and the information used by the time-series models.

After identifying these bridge variables, we can start working on creating a schema for the knowledge graph. The goal of this step is to structure the schema around these variables so that these variables can transfer as much information as possible from the knowledge graph to the time-series models. For variables, like ”date” this is easy as most of the data should already be dated. For variables that cannot be directly connected to the bridge variables, we have to look into the number of hops necessary to reach these bridge variables. The greater the number of hops, the lesser the chances of information being transferred from the embeddings to the time-series models and vice versa. After integrating the knowledge graph with the time-series models, future research can look into the impact of the number of hops on the performance of the time-series models.

The knowledge graph also needs to be temporal i.e. changing with time, meaning as new data comes in the knowledge graph is updated resulting in new embeddings. The reason for doing this is to prevent any data leakage from using future sources of data in the knowledge graph, which can overfit the time-series model. In practice, this is done by first creating the knowledge graph till a specific date. Then in the next step, we add data for the next date and repeat. When using temporal graphs we can add an auxiliary unsupervised task of predicting the graph structure in the next step, along with any existing supervised task, as multi-task learning has been shown to benefit computer vision and natural language processing tasks, and we can incorporate these into the graph domain also.

1.2. Creating Embeddings

In the previous section, we defined a schema and created a temporal knowledge graph. This step can be done in Neo4j (or Scalation in the future) which provides a tested knowledge graph database. Using Neo4j, we can define the schema and load the necessary data and then use one of the built-in methods (like GraphSage) to train embeddings which can be passed to the time-series models.

Specifically, every node is associated with an N-dimensional vector and every edge is associated with an R-dimensional vector, which will be learned by a model. Models like GraphSage, ComplEx, and TransE will take these vectors as inputs and modify them to minimize the specified loss function. Internally, these models build upon the message-passing algorithm where a node first collects information from all the neighboring nodes and then passes it further.

The model aims to learn embeddings that best capture the graph structure, meaning captures the maximum amount of information from the knowledge graph to the embedding vectors.

An example of the learned embeddings is shown in Fig. 4 where we visualize the 2D embeddings[1]. Using the figure, we can look at individual states and the states that are closer to each other share similar COVID-19 patterns as learned from the knowledge graph. The embeddings projection cannot be used to verify the quality of the knowledge graph, as it is not possible to differentiate from random embeddings. So the only way to evaluate the quality of the embeddings obtained from the knowledge graph is to pass them to the time series model and see if the metrics improve.

1.3. Integrating with time-series models

Bridge variables are used to integrate the knowledge graph and the input for the time-series models. Since bridge variables are generally categorical like date, and state which are usually initialized as random/one-hot-encoded embedding vectors, we can directly replace these with the embeddings learned from the knowledge graph and pass them as input to the time-series models. To reduce the noise in the embeddings, the embeddings can be projected to a lower latent space using UMAP, T-SNE, or PCA.

[1] https://projector.tensorflow.org/ is used to get the visualization

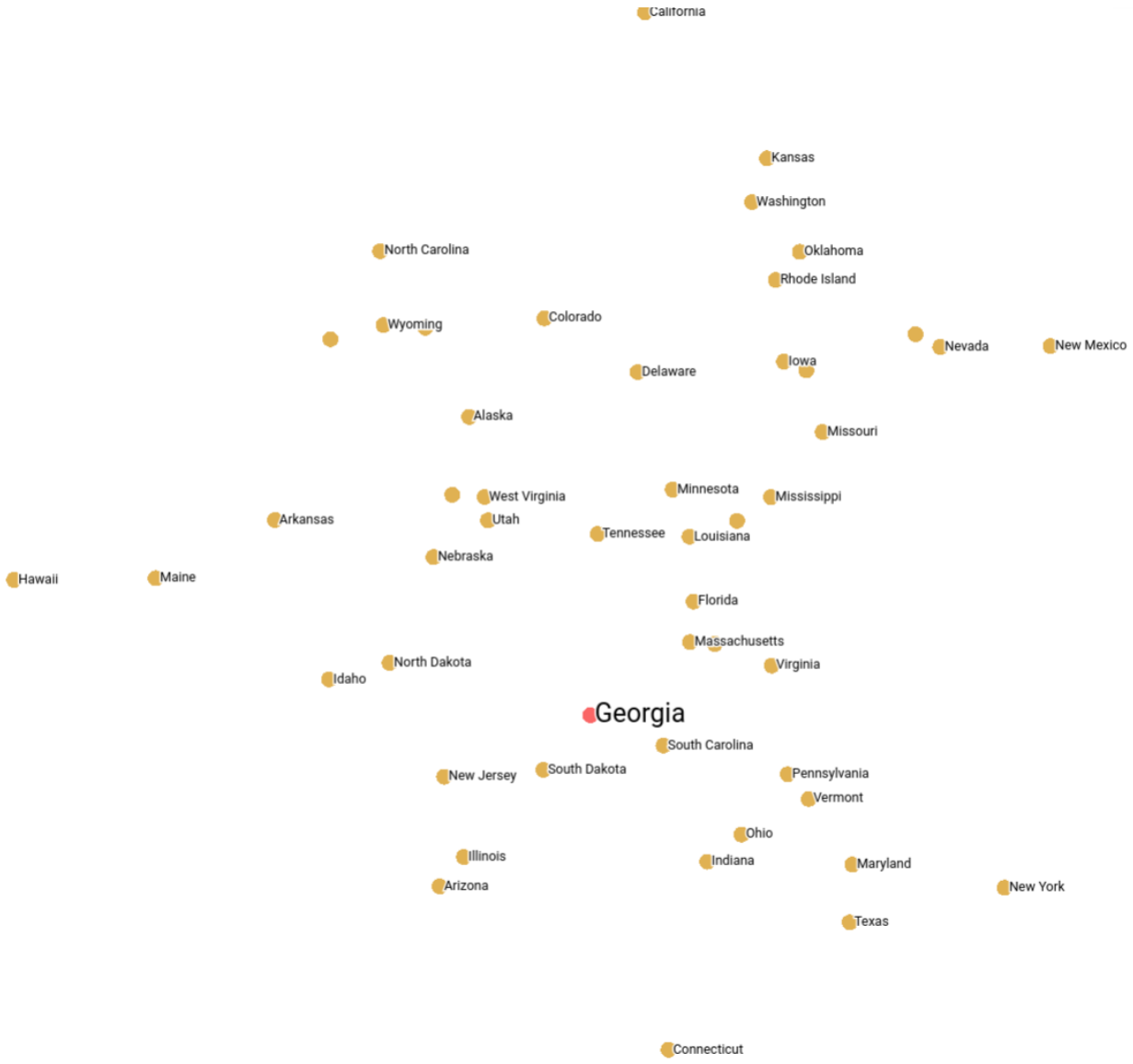


Figure 4. 2D projection of the state embeddings obtained from the knowledge graph using PCA. States that are closer to each other share similar COVID-19 patterns as learned by the models from the knowledge graph. To verify the quality of the embeddings we cannot simply look at the figure, because it is not possible to differentiate these embeddings from random embeddings. And thus we have to pass the embeddings to the time series model and see if the metrics improve.